ANNUAL PROGRESS REPORT
05/01/2007 - 04/30/2008

NASA 05-AISR05-0143

Bayesian Source Separation for Astrophysical Spectra: Application to PAHS

Kevin H. Knuth (PI)
kknuth@albany.edu (518) 442-4653
Department of Physics, PH 211, University at Albany, Albany NY 12222
http://knuthlab.rit.albany.edu/

General Summary

We propose to apply Bayesian source separation techniques to a very important subset of astrophysical spectra, the emission spectra of Polycyclic Aromatic Hydrocarbons (PAHs). The PAH species, ubiquitous in our Galaxy and in external galaxies, are one of the most important molecular species in the pathway to life. Their spectra are the result of the blending of the complex spectra of individual PAH species, possibly at different temperatures, and to date they have defied all attempts to unscramble them into individual contributors. We propose to apply Bayesian source separation for the first time in order to quantitatively determine which of the many hundreds of individual PAH species contribute most significantly to the combined emission spectra of PAHs. Our research will lead to powerful new tools for decomposing, analyzing and characterizing spectra of these critical species. Furthermore, these tools will be readily generalizable to a wide array of spectral analysis problems ranging across astrophysics, Earth science, and biology.

Summary of Our Efforts

Our second year builds on the successes of the first year, which included demonstrating feasibility of the approach with 80-100 mixed PAHs using non-negative least-squares, and the development of two algorithms to characterize the background spectra. This year, we have focused on extending the nested sampling algorithm to perform PAH estimation. This has proven to be extremely challenging for a number of reasons, which we will describe below in the section titled "Schedule Status, Delays/Problems Experienced, and Corrective Actions Plan".

At this stage we have extended the simple PAH mixture model, which consists of a set of estimated concentrations for the PAHs to a more complex model, which introduces, in addition to the set of concentrations, a set of discrete indicator parameters that indicate the presence or absence of a PAH independent of concentration. The purpose of this extension is to enable us to infer the probability that a PAH is either present or absent in the event that the PAH concentration is small. Without these indicator parameters, it is impossible to interpret an analysis that results in a PAH having an extremely small concentration. Is it the case that the PAH is present with high probability, but has a miniscule concentration? Or is it the case that the PAH is in actuality absent and that fact that the estimated concentration is small implies that the PAH is absent? The presence

indicator parameters enable us to compute the probability of presence or absence independent of concentration. Implementation of this aspect of the algorithm using nested sampling has proven to be extremely difficult. We need to ensure adequate exploration of the Markov chain, while simultaneously ensuring that computations are not wasted due to an inordinate number of rejections. This has required that we give careful consideration to the moves that are made by the Markov chain so that the difficulties associated with model-order transitions are minimized. At this point we are continuing to polish these aspects of the algorithm.

One of Knuth's (PI) graduate students, Haley Maunu, spent the summer at NASA Ames Research Center working with Duane Carbon (co-I). Haley worked to develop and test nested sampling code on spline models while working with Duane to evaluate the spectral data from a wide variety of PAHs.

The Ames team (Allamandola, Bauschlicher, Cami and Peeters) continued work to build their database of PAH spectra. This database will be essential for our future algorithms which will test for the presence of these PAH species in real datasets.

Personnel, Progress Description, and Summary of Accomplishments
In May of 2007, we hired Dr. Deniz Gencaga, who is an expert in source separation applied to astrophysical spectra, as a Senior Research Support Specialist. Deniz has been essential in working with the nested sampling algorithm and implementing the discrete indicator parameters. Deniz Gencaga has also developed measures of PAH similarity, which are used by our sampling algorithm to facilitate exploration.

In the summer of 2007, we hired three physics graduate students Nabin Malakar, Man Kit Tse, and Haley Maunu to work on developing MatLab code for PAH spectrum classification. Man Kit Tse completed his efforts on the spectral background estimation algorithm, which was based on a Mixture of Gaussians model and implemented with the nested sampling algorithm. This work was presented at the MaxEnt 2007 Workshop, Saratoga Springs NY in July 2007, and the IEEE Statistical Signal Processing Workshop, Madison WI in August 2007. Nabin Malakar worked to explore sampling methods and source separation. Haley Maunu, spent the summer at NASA Ames Research Center working with Duane Carbon (co-I). She worked to develop and test nested sampling code on spline models while working with Duane to evaluate the spectral data from a wide variety of PAHs.

Undergraduate physics majors Christopher Stiles and David Idell worked with the PI for senior research credit to develop planetary nebula modeling code for use on our laboratory's 16-node Beowulf cluster. These individuals have made incredible progress and we now have modeling and inference code running on our cluster. They are currently working to characterize planetary nebula by relying on data from the Hubble Space Telescope. These efforts are related to this PAH project as planetary nebulae are rich in PAHs and serve as the environment in which the astrobiochemistry we are investigating takes place.

Duane Carbon at NASA Ames has continued to investigate the behavior of the non-negative least-squares algorithm, and to work with the PAH group to build the PAH database that is central to our efforts. Duane worked to mentor Haley Maunu for the summer of 2007 and together they studied both the PAH and background spectra in detail.

Kevin Knuth has worked to oversee this project as well as guiding the students on the developments using the Nested Sampling algorithm and integrating the algorithmic results across the team. He has also given several talks on the topic as well as writing abstracts and papers with Duane Carbon and Deniz Gencaga. He has worked with Undergraduates, Chris Styles and David Idell, to integrate our PAH estimation efforts with past efforts in characterizing the environments of planetary nebulae which are the sources of many of the PAH spectra we are studying. Last year, we reported advances in a Bayesian histogramming technique that we were polishing up for submission into the AISRP Code and Algorithm Library. These efforts were extended and the algorithm and research paper are now complete. The paper describing this technique will be submitted to *Phys Rev E.*

Schedule Status, Delays/Problems Experienced, and Corrective Actions Plan
Estimation of the discrete indicator parameters in the PAH mixture model using nested sampling has proven to be extremely difficult. We need to ensure adequate exploration of the Markov chain, while simultaneously ensuring that computations are not wasted due to an inordinate number of rejections. This has required that we give careful consideration to the moves that are made by the Markov chain so that the difficulties associated with model-order transitions are minimized. We have also found several problems with the way that this algorithm was originally implemented by Skilling that caused difficulties with our specific problem. These problems now have been fixed.

Work Plan for next Reporting Period
We are currently working to fine tune the specific moves that are used by our nested sampling algorithms. For the remainder of this school year and into the summer of 2008, we will be working to test the PAH estimation code on both synthetic mixtures and real data acquired by the Infrared Space Observatory (ISO) and the Spitzer Space Telescope. We will also work to identify classes of PAHs that are difficult to separate (some of this has already been performed by Duane Carbon). By June 2008, we expect to be very familiar with the algorithms ability to perform PAH identifiability. As we move into the next year of funding, we will rely on the success of our techniques to begin to deal with more subtle problems such as sets of PAHs at distinct temperatures.

Publications and Presentations

1. Knuth, K.H., Tse M.K., Choinsky J., Maunu H.A, Carbon D.F. 2007. Bayesian source separation applied to identifying complex organic molecules in space. Proceedings of the IEEE Statistical Signal Processing Workshop, Madison WI, August 2007.

2. Tse M.K, Choinsky J., Carbon D.F. and Knuth K.H. 2007. Estimating background spectra. In: Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Saratoga Springs, NY, USA, 2007, K.H. Knuth, A. Caticha, J.L. Center, A. Giffin, C.C. Rodriguez (eds.), AIP Conference Proceedings 954, American Institute of Physics, Melville NY, 322-327.

3. Gençağa D., Carbon D.F. Knuth K.H. 2008. Characterization of interstellar organic molecules. Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2008), São Paulo, Brazil, July 2008.

In Preparation

4. Knuth K.H. Optimal data-based binning for histograms. *Phys Rev E*

5. Gençağa D., Carbon D.F. Knuth K.H. 2008. Source separation applied to organic molecules in interstellar space. *Phys Rev E*

Synergistic Activities

Below I list our synergistic activities. The PI has chaired several conference sessions and served on technical committees for signal processing conferences where the new techniques we are investigating for this grant are supported, presented and discussed.

1. Edited Special Journal Issue on Source Separation:
Digital Signal Processing Special Issue on Bayesian Source Separation, Volume 17, Issue 5, Pages 855-994. Guest Editors: Kevin H. Knuth and Ercan Kuruoglu

2. Knuth K.H. (Editor), Caticha A. (Editor), Center J.L. (Editor), Giffin A. (Editor), Rodríguez C.C. (Editor). *Bayesian Inference and Maximum Entropy Methods in Science and Engineering, Saratoga Springs, NY, USA, 2007*, AIP Conference Proceedings 954, American Institute of Physics, Melville NY.

2007   Science Session Organizer and Chair for Data Mining in Aeronautics, Science and Exploration Systems (DMASES) 2007 Conference, June 25-27, 2007, Mountain View CA.

2007   Organizer for the 27th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt) 2007, July 8-13, 2007, Saratoga Springs NY.

2007   Technical Committee Member for the European Signal Processing Conference (EUSPICO) 2007, Sept. 3-7, 2007, Poznan, Poland.

2007   Reviewer for 7th International Conference on Independent Component Analysis and Signal Separation ICA 2007, Sept. 9-12, 2007, London UK.

2008   Journal on Advances in Signal Processing (JASP) Best Paper Award Sub-Committee Member.